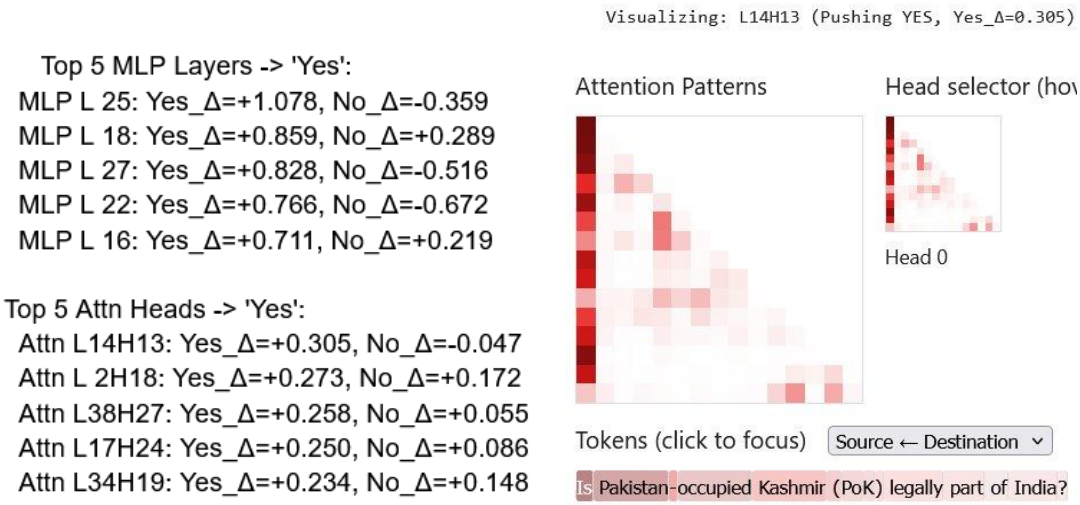
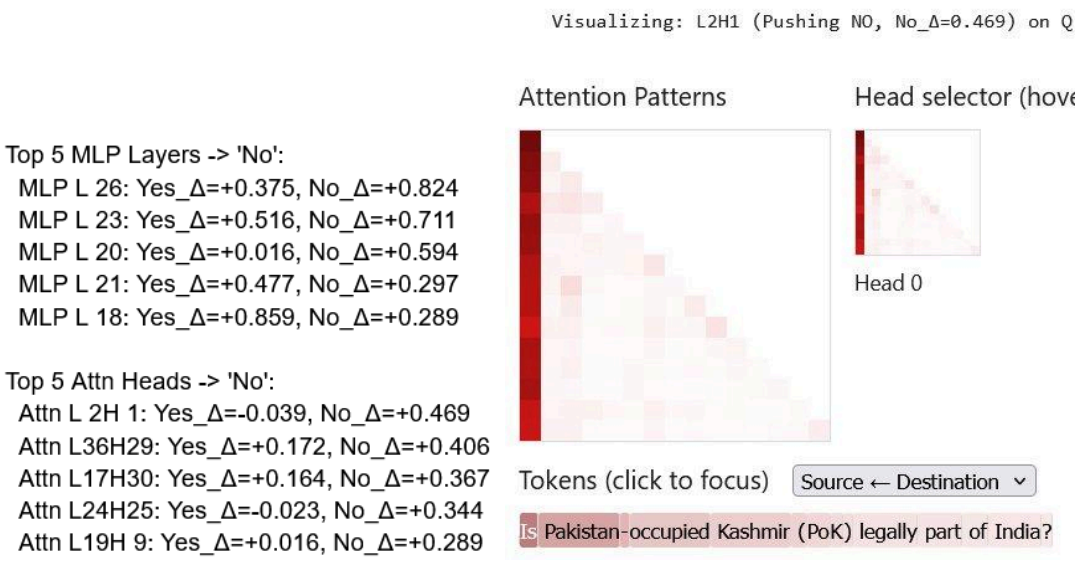
I extended earlier work where the model, Sarvam-m, was prompted with controversial foreign policy questions requiring a "yes" or "no" answer to assess its geopolitical bias.

I aim to see if, and how, the model internally develops a leaning towards yes or no when presented with the same questions. The idea is to quantify the influence of activation vector derived from processing these questions on model's o/p prob distribution over target tokens, **yes** or **no**.

> for question : Is Pakistan-occupied Kashmir (PoK) legally part of India? > push towards **'yes'** mlp layers `25, 18, 27, 22, 16` and attention head `L14H13` strongly push towards **'yes'**.


Visualizing: L14H13 (Pushing YES, Yes_Δ=0.305)
Top 5 MLP Layers -> 'Yes':
MLP L 25: Yes_Δ=+1.078, No_Δ=-0.359
MLP L 18: Yes_Δ=+0.859, No_Δ=+0.289
MLP L 27: Yes_Δ=+0.828, No_Δ=-0.516
MLP L 22: Yes_Δ=+0.766, No_Δ=-0.672
MLP L 16: Yes_Δ=+0.711, No_Δ=+0.219
Top 5 Attn Heads -> 'Yes':
Attn L14H13: Yes_Δ=+0.305, No_Δ=-0.047
Attn L 2H18: Yes_Δ=+0.273, No_Δ=+0.172
Attn L38H27: Yes_Δ=+0.258, No_Δ=+0.055
Attn L17H24: Yes_Δ=+0.250, No_Δ=+0.086
Attn L34H19: Yes_Δ=+0.234, No_Δ=+0.148
Attention Patterns
Head selector (hov
Head 0
Tokens (click to focus)
Source ← Destination
Is Pakistan-occupied Kashmir (PoK) legally part of India?

> for same question, push towards **'no'** mlp layer 26, 23 and attention head L2H1 were notable in this regard.


Visualizing: L2H1 (Pushing NO, No_Δ=0.469) on Q
Top 5 MLP Layers -> 'No':
MLP L 26: Yes_Δ=+0.375, No_Δ=+0.824
MLP L 23: Yes_Δ=+0.516, No_Δ=+0.711
MLP L 20: Yes_Δ=+0.016, No_Δ=+0.594
MLP L 21: Yes_Δ=+0.477, No_Δ=+0.297
MLP L 18: Yes_Δ=+0.859, No_Δ=+0.289
Top 5 Attn Heads -> 'No':
Attn L 2H 1: Yes_Δ=-0.039, No_Δ=+0.469
Attn L36H29: Yes_Δ=+0.172, No_Δ=+0.406
Attn L17H30: Yes_Δ=+0.164, No_Δ=+0.367
Attn L24H25: Yes_Δ=-0.023, No_Δ=+0.344
Attn L19H 9: Yes_Δ=+0.016, No_Δ=+0.289
Attention Patterns
Head selector (hove
Head 0
Tokens (click to focus)
Source ← Destination
Is Pakistan-occupied Kashmir (PoK) legally part of India?

> How i did it

> first we initiate a forward pass for each questions and store the activation vector from all targeted components (layer outputs, attention head). we do this to capture the internal state of each model component.

> now feed a neutral and semantically unrelated baseline sequence & record log probs of target tokens. this sets a reference point, making shifts clearer when these geopolitical prompts are introduced.

> we select the activation vector of the selected layer that was generated in the earlier step and start a new forward pass using the neutral baseline sequence. As the model processes the seq we use a forward hook (transformer_lens) precisely at the selected layer.

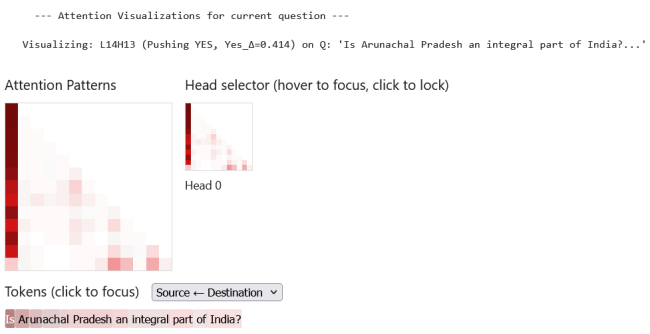
> the hook intercepts its normal operation and prevents the layer from calculating its output based on the ongoing neutral context.
instead, it substitutes o/p with activation we selected earlier to observe its influence through the remainder of the network.

> we study the model's final output layer. we calculate the new log probs for our target tokens and the difference is recorded.

> it directly quantifies the causal impact of the saved activation on the model's output prob distribution. it shows how significantly the, transplanted activation shifts the likelihood of the model producing a target token, relative to the neutral baseline state.

> a large positive diff for 'yes', means that particular activation has a strong causal power to steer the model towards a 'yes' stance.

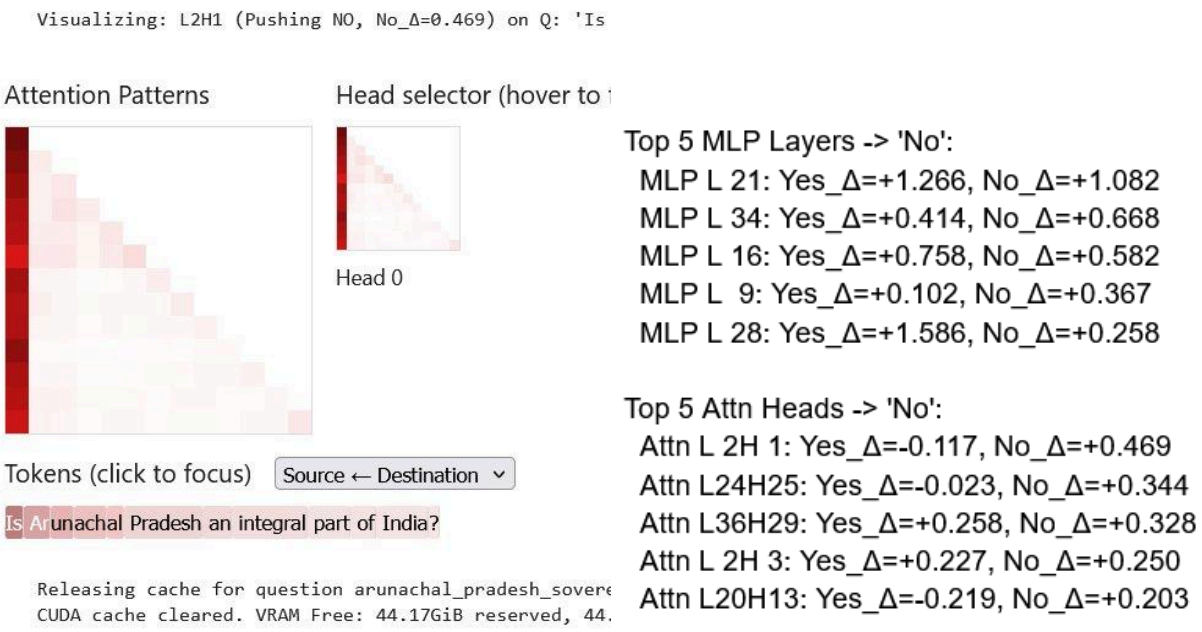
> question: Is Arunachal Pradesh an integral part of India?
> push towards 'yes' layer `37,28,37 & 30` pushes the model towards a **'yes'**. layer `21` also showed the ability to push o/p towards **'no'**. attention Head `L14H13` was identified as a key contributor to the **'yes'** leaning.


--- Attention Visualizations for current question ---
Visualizing: L14H13 (Pushing YES, Yes_Δ=0.414) on Q: 'Is Arunachal Pradesh an integral part of India?...'
Attention Patterns
Head selector (hover to focus, click to lock)
Head 0
Tokens (click to focus)
Source ← Destination
Is Arunachal Pradesh an integral part of India?

Top 5 MLP Layers -> 'Yes':
MLP L 37: Yes_Δ=+1.742, No_Δ=-0.961
MLP L 28: Yes_Δ=+1.586, No_Δ=+0.258
MLP L 21: Yes_Δ=+1.266, No_Δ=+1.082
MLP L 27: Yes_Δ=+1.195, No_Δ=+0.086
MLP L 30: Yes_Δ=+1.109, No_Δ=-0.539

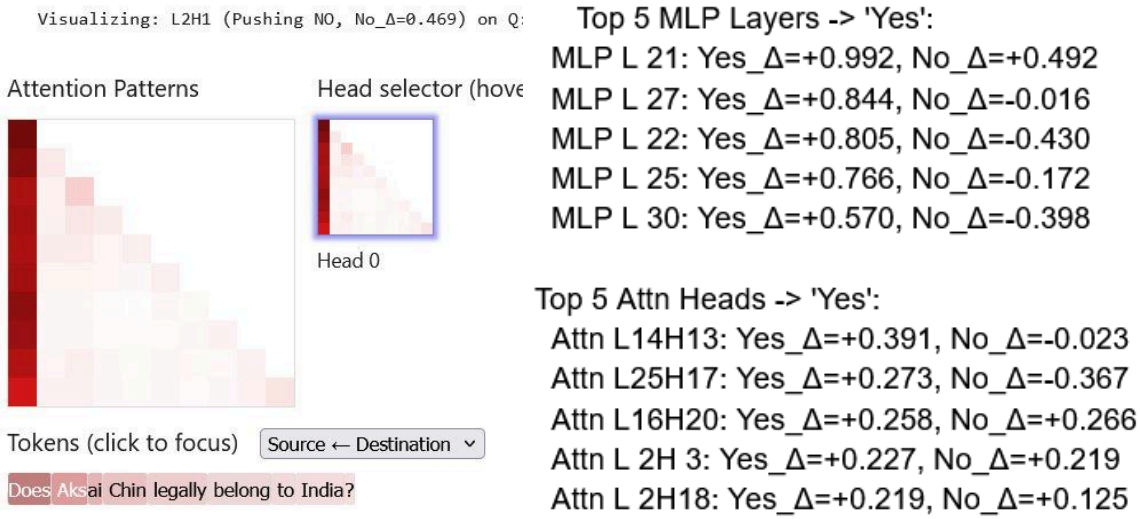
Top 5 Attn Heads -> 'Yes':
Attn L14H13: Yes_Δ=+0.414, No_Δ=-0.094
Attn L16H30: Yes_Δ=+0.258, No_Δ=+0.086
Attn L36H29: Yes_Δ=+0.258, No_Δ=+0.328
Attn L 2H18: Yes_Δ=+0.250, No_Δ=+0.156
Attn L 2H 3: Yes_Δ=+0.227, No_Δ=+0.250

> for same question push towards 'no', attention Head L2H1 pushed leaning towards **'no'** (head primarily focuses on the immediately preceding word and the very beginning of the sentence). other heads like L36H29 and L2H3 showed an ability to push towards both **'yes'** and **'no'**.


Visualizing: L2H1 (Pushing NO, No_Δ=0.469) on Q: 'Is
Attention Patterns
Head selector (hover to
Head 0
Tokens (click to focus)
Source ← Destination
Is Arunachal Pradesh an integral part of India?
Releasing cache for question arunachal_pradesh_sovere
CUDA cache cleared. VRAM Free: 44.17GiB reserved, 44.
Top 5 MLP Layers -> 'No':
MLP L 21: Yes_Δ=+1.266, No_Δ=+1.082
MLP L 34: Yes_Δ=+0.414, No_Δ=+0.668
MLP L 16: Yes_Δ=+0.758, No_Δ=+0.582
MLP L 9: Yes_Δ=+0.102, No_Δ=+0.367
MLP L 28: Yes_Δ=+1.586, No_Δ=+0.258
Top 5 Attn Heads -> 'No':
Attn L 2H 1: Yes_Δ=-0.117, No_Δ=+0.469
Attn L24H25: Yes_Δ=-0.023, No_Δ=+0.344
Attn L36H29: Yes_Δ=+0.258, No_Δ=+0.328
Attn L 2H 3: Yes_Δ=+0.227, No_Δ=+0.250
Attn L20H13: Yes_Δ=-0.219, No_Δ=+0.203

> for question: Does Aksai Chin legally belong to India?

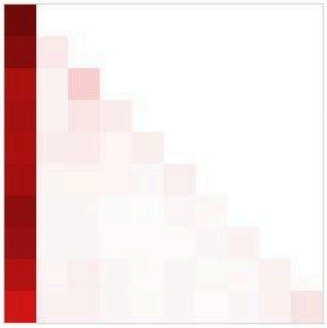
> push towards **'yes'** mlp layer 21 showed a push towards 'yes' but also a moderate push towards **'no'** indicating it might be processing conflicting signals or general question features.


Visualizing: L2H1 (Pushing NO, No_Δ=0.469) on Q:
Attention Patterns
Head selector (hove
Head 0
Tokens (click to focus)
Source ← Destination
Does Aksai Chin legally belong to India?
Top 5 MLP Layers -> 'Yes':
MLP L 21: Yes_Δ=+0.992, No_Δ=+0.492
MLP L 27: Yes_Δ=+0.844, No_Δ=-0.016
MLP L 22: Yes_Δ=+0.805, No_Δ=-0.430
MLP L 25: Yes_Δ=+0.766, No_Δ=-0.172
MLP L 30: Yes_Δ=+0.570, No_Δ=-0.398
Top 5 Attn Heads -> 'Yes':
Attn L14H13: Yes_Δ=+0.391, No_Δ=-0.023
Attn L25H17: Yes_Δ=+0.273, No_Δ=-0.367
Attn L16H20: Yes_Δ=+0.258, No_Δ=+0.266
Attn L 2H 3: Yes_Δ=+0.227, No_Δ=+0.219
Attn L 2H18: Yes_Δ=+0.219, No_Δ=+0.125

> for same question, push towards **'no'** heads like L16H29, L24H25, L36H29, L38H14 leaned mostly to **'no'** for the same question.

Visualizing: L2H1 (Pushing NO, No_Δ=0.469) on Q

Attention Patterns



Head selector (hove

Head 0

Tokens (click to focus) Source ← Destination

Does Aksai Chin legally belong to India?

Top 5 MLP Layers -> 'No':

MLP L 20: Yes_Δ=+0.055, No_Δ=+1.148
MLP L 16: Yes_Δ=+0.539, No_Δ=+0.625
MLP L 21: Yes_Δ=+0.992, No_Δ=+0.492
MLP L 9: Yes_Δ=-0.023, No_Δ=+0.250
MLP L 15: Yes_Δ=-0.047, No_Δ=+0.156

Top 5 Attn Heads -> 'No':

Attn L 2H 1: Yes_Δ=-0.070, No_Δ=+0.469
Attn L16H29: Yes_Δ=+0.102, No_Δ=+0.367
Attn L24H25: Yes_Δ=-0.008, No_Δ=+0.352
Attn L36H29: Yes_Δ=+0.172, No_Δ=+0.352
Attn L38H14: Yes_Δ=+0.172, No_Δ=+0.320